CLOZE TESTING: Analysis and Problems

Kenji OHTOMO

The cloze procedure, originated by Wilson L. Taylor (1953), has received considerable attention in the field of testing English as a second or foreign language. Donald K. Darnell (1968) and John W. Oller (1973a), as well as many other researchers, have acknowledged its importance. Previous studies have provided rather convincing support for the value of cloze tests. Some specialists, however, have begun questioning the principle and methods of the cloze test as a formal instrument of measuring the proficiency of English as a second or foreign language. The writer attempts to present in this paper an overview of the theory underlying the cloze test, major findings concerning the techniques of conducting the test, and finally to point out certain problems of the test.

I. Introduction

It is quite appropriate to state that the principle and methods of testing English as a foreign language tend to follow those of teaching and learning English. In the field of testing the learner's English proficiency as in the teaching and learning of English, there has been considerable shifting from one approach to another. We are, however, able to divide the series of the movements into three periods as discussed by Bernard Spolsky (1978; v-x).

Spolsky has assigned the following names to each period: 1) the pre-scientific period, 2) the psychometric-structuralist period and 3) the integrative-sociolinguistic period. Let us first examine the characteristics of each of these.

1.1. Pre-Scientific Period

During this period, the first thing we can say concerning the test of English is that little consideration was made of statistical matters such as validity, reliability or even practicality which are the fundamental characteristics of a good test. As Spolsky (1978; v) mentioned, "...if a person knows how to teach, it is to be assumed that he can judge the proficiency of his students."

In our country, this period is the one when the main purpose of teaching English was to translate English into Japanese or vice versa. Grammar translation was the principle teaching method of English during this period. Most of the item types of the test were the translation into or from the English language. Free composition was one of the representative item types, where the teacher just gave the title or the topic such as "my hobby," "my family," "The most exciting experience I have ever had" and so on. Such methods were naturally very subjective.

Due to the nature of the grammar translation method adopted for teaching, most of the test was conducted using paper and pencil techniques. There was very little trial in the evaluation of listening and speaking.

1.2. Psychometric-Structuralist Period

The psychometric-structuralist period may be characterized by, first of all, "objective" measures using various statistical techniques and, second, the notion of structural linguistics. This may be represented by the approach to foreign language testing developed by Robert Lado (1961) and his colleagues.

Sometimes labelled as "the discrete point approach," it can be summarized as follows:

The testing points should be based on the trouble spots found in the contrastive study of the mother tongue and the target language. Moreover, measurement should be conducted at the different levels

of syntax, morphology, and phonology, concentrating on the specific items found in the contrastive study. In other words, knowledge of these items constitutes knowing a particular language. According to their way of thinking, language ability is thought to be divisible. It follows that the goal of studying the language materials in question is to be achieved by developing skills at the different levels of syntax, morphology and phonology. If we use materials developed according to that approach and wish to evaluate what students have studied after a certain period of time with a certain type of teaching method, then, achievement tests based on this approach might be of great use. However, in general proficiency tests, where information is needed for determining language capability as a result of cummulative learning experience regardless of the materials and methods used, this approach is insufficient and inadequate. Rebecca M. Valette (1967), David P. Harris (1969), and John L. D. Clark (1972) are all, even though their emphases vary, representative of the psychometric-structuralist period.

1.3. Integrative-sociolinguistic Period

Even though there seemed to be a large number of researchers who insisted on the importance of the integrative-sociolinguistic phase of the foreign language test, it should be noted that John B. Carroll (1961) was the first to criticize the discrete point approach. Carroll insisted on the importance of the total communicative effect of the utterance rather than its discrete linguistic components. Carroll (1968; 57) is unique in the sense that he pointed out the integrative elements to be measured which were not noticed by Robert Lado and his following. They are, for example, 1) speed of response, 2) diversity of response, 3) complexity of information process, and 4) awareness of linguistic competence.

It should also be noticed that Clark (1972; 119), though he is classified as one of the researchers in the psychometric-structuralist

period, has divided language ability into two: linguistic ability and communicative proficiency. Linguistic ability is defined in such terms as "...accuracy of pronunciation, range of vocabulary, accuracy and extent of grammatical control, and so forth...," while communicative proficiency "...ability to get a message across to an interlocutor with a specified ease and effect." What is of further interest to the writer is that "...between these two extremes, linguistic proficiency *per se* and the ability to communicate readily and effectively in real-life situations have a tenuous correlation."

It is clearly found that the works by Bernard Spolsky, *et al.* (1968), Donald K. Darnell (1968), Leon A. Jakobovits (1969), John Oller (1973, b) are surely indebted to the insight deepened by John B. Carroll.

The basic principle of the non-discrete point approach, integrative approach, or integrative sociolinguistic approach is that, first of all, the knowledge of a language is more than just the sum of discrete parts, and that, secondly, we need to add a strong functional dimension to the language testing. The assumption by Spolsky *et al.* (1968; 81), as one of the typical principles of the integrative-sociolinguistic approach, that "There is such a factor as overall proficiency in second language, and it may be measured by testing a subject's ability to send and receive messages under varying conditions of distortion of the conducting medium" is thought to be of genuine value.

In order to put this new principle into practice, a large number of language tests have been developed. Among these are "Clozentropy Procedure" by Donald K. Darnell (1968), "Noise Test" by Bernard Spolsky *et al.* (1968), "Productive Communication Testing" by John A. Upshur (1973), "The Bilingual Syntax Measure" by Marina K. Burt, Heidi C. Dulay, and Eduardo Hernandez-Chavez (1973), "The Foreign Service Institute Oral Interview Test" developed by The Foreign Service Institute, "Ilyin Oral Interview" by

Donna Ilyin (1976) and so on. "Dictation" and "Cloze Test" were re-evaluated as useful instruments in order to measure the overall proficiency. John W. Oller and a lot of other scholars have been doing their best in order to improve the testing methods.

The writer hopes that with this very brief historical sketch, the the reader will be able to locate "cloze testing" in the stream of the history of foreign language testing.

II. Analysis

2.1. Underlying Theory

2.1.1. Definition and Procedure

Wilson L. Taylor (1953; 416) defines the cloze procedure as "...a method of intercepting a message from a 'transmitter,' mutilating its language patterns by deleting parts, and so administering it to 'receivers' that their attempts to make the patterns whole again potentially yield a considerable number of cloze units" The cloze procedure has antecedents in Gestalt psychology and in the common sentence completion technique. It was originally developed as an index of readability and was also defined by Taylor (1954; 3) as "...a psychological tool for gauging the degree of total correspondence between (1) the encoding habits of transmitters and (2) the decoding habits of receivers." In Taylor's (1956; 48) application to readability measurement, every fifth word was deleted because he found it made optimum use of the sampled material and allowed all sort of words to be represented according to the proportion of their occurrence.

Thus the procedure of the cloze test is quite simple. Filling the blanks by guessing the missing words in the text is, according to Taylor's notion; a special kind of "closure," from which the term "cloze" originates. It is believed that Taylor is responsible for coining the word "cloze" rather than the word "close." In the practical method, you delete every *n*th word from a passage picked

up, and ask the examinees to supply the missing words in the blanks provided. It is generally believed that a person who is either a native speaker of the language to be tested, English for example, or a non-native speaker of the language who is reasonably proficient should be able to fill in the blanks if given a context. What is the difference between the traditional 'fill-in-blank' test over single sentence and 'cloze' test? Taylor (1953; 417) points out the basic differences between them as follows: first, "cloze procedure deals with contextually interrelated series of blanks, not isolated ones"; and second, "...the cloze procedure does not deal directly with specific meaning. Instead it repeatedly samples the likeness between the language patterns used by the writer to express what he meant and those possibly different patterns which represent readers' guess at what they *think* he meant." This can be understood by just the fact that he included a subsection entitled, 'Not a Sentence-Completion Test.'

2.1.2. Internalized Grammar of Expectancy

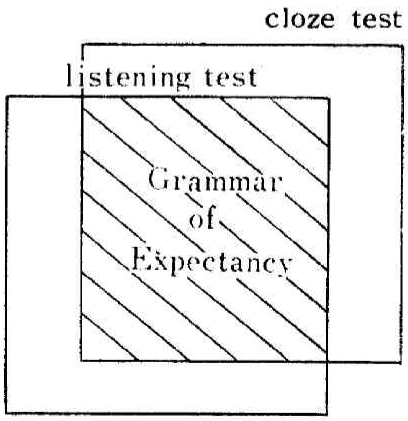
One of the most exciting findings in the study of cloze tests in foreign language testing is that the results of the scores of cloze tests correlates usually very highly with listening comprehension tests. Cloze tests are directly connected with written language, as do listening comprehension tests with spoken language. Why do the results of the written and spoken language tests correlate so high? If you can find, and formulate the answer to this question, then you will be able to explain Oller's hypothesis of the cloze tests.

It has been agreed that cloze tests provide a highly integrative and useful device in assuring foreign students' general proficiency. The point we have to make clear is the meaning of general proficiency. As we have already mentioned before, the discrete point approach is thought to be a direct reflection of the notion of the structural linguistics with the strong emphasis of reliability of the test. This basic idea leads to the fact that if you get across

thousands of structural items, you will have taught the language. The main reaction to this assumption, however, is that whether or not thousands of structural patterns isolated from meaningful contexts of communication constitute language competence. In the same fashion, the main problem is whether or not thousands of these discrete point items constitute adequate testing of language competence. It is generally believed by the people who are against discrete point approach that thousands of discrete point items do not constitute an appropriate measurement of language competence.

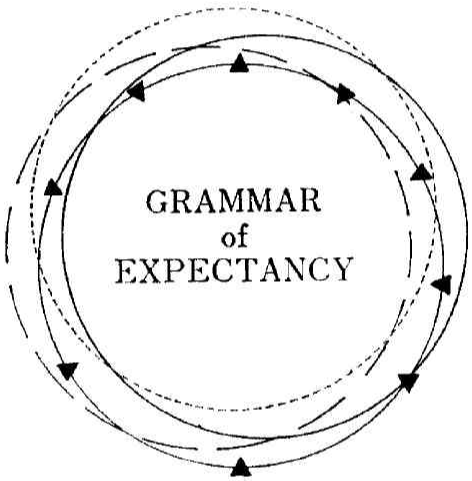
The logic behind cloze testing is, according to Oller (1979, 32), that "...language testing is primarily a task of assessing the efficiency of the pragmatic expectancy grammar the learner is in the process of constructing." A pragmatic expectancy grammar is defined, as Oller (1979; 34) states, "...as a psychologically real system that sequentially orders linguistic elements in time and in relation to extralinguistic contexts in the meaningful ways."

It is quite interesting to note that the results of the cloze test which is concerned exclusively with written language, should correlate so highly with spoken language test results, such as dictation and listening comprehension tests. The reason for the existence of such a high correlation between them may be explained by "expectancy grammar." It has been assumed that the learner's internalized grammar of expectancy is the central component of language



competence. If the cloze tests correlate with listening comprehension tests at .90 level, it means that roughly 81% of the variance on the test is common variance. The preceding figure shows variance overlap on integrative test as an indication of an underlying grammar of expectancy.

It has often been found that integrative, or non-discrete point tests tend to have strong intercorrelation with each other. It can be illustrated in the following chart, as mentioned by Oller (1978; 54):



2.1.3. Unitary Competence Hypothesis

What we may be able to hypothesize, if we follow Oller's notion of grammar of expectancy, is that we will be able to deny that language skill might be divided up into several components, skills, or aspects. This hypothesis is often called 'indivisibility hypothesis,' or 'unitary competence hypothesis.' The hypothesis that discrete point testers have suggested is often called 'divisibility hypothesis.' There may be a third hypothesis in connection with the view of foreign language proficiency; that is, one that goes between the two extremes mentioned above. 'The partial divisibility hypothesis' may

here be the appropriate label. In that hypothesis, it is thought that there will be a large amount of components common to all of the test, plus a small amount of components shared by only some of the tests.

With the unitary competence hypothesis, it is quite possible to reason the assumption that "...valid language tests will intercorrelate at very high levels, regardless of modality, format, etc." as mentioned by Oller (1976; 152). The assumption means that if a test is measuring what it is supposed to measure, it will correlate at a high level with other valid tests whatever the modality or format. This leads to the fact that if the tests are valid, the test results of listening, speaking, reading, writing, and any othe integrative test correlate with each other at a very high degree of level. In other words, expectancy grammar will be measured by any test if the test is valid. It is for this reason that Oller tends to support the unitary competence hypothesis. Most of the studies in Oller, J. W. and Perkins, K. eds. (1980) seem to support the unitary competence hypothesis. Among those studies are Oller and Hinofotis (1980; 13–23); Scholz, Hendricks, Saurling, Johnson and Vandenburg (1980; 24–33), etc.

2.2. Major Findings

Now, what are the findings of cloze tests? Is a cloze test an automatically valid procedure? The writer must say 'no.' It seems a little misleading if we regard the cloze test as an automatically valid procedure which results in an universally valid test. There must be clear distinction between the assumption and real findings on what is true or not in the cloze test. In this part of the paper, the writer would like to summarize what has been found in the study during the past so that we would be able to pick up the items that remain unsolved. Among the techniques of giving the cloze test, let us see what has been found in 1. scoring methods,

2. deleting methods, 3. length of the text, 4. difficulty level of the passage, and 5. content of the passage for the use of the test.

2.2.1. Scoring Methods

There seems to be four scoring methods in the cloze test. They are a) exact-word scoring method, b) acceptable-word scoring method, c) clozentropy scoring method and d) multiple-choice scoring method.

The exact-word scoring method is the one that gives a point when the student fills the blank with only the words deleted from the text. The second method, the acceptable-word scoring method, is the one that gives a point for any grammatically and contextually appropriate response. The third one is the most complex method. This method was developed in 1968 by Darnell and modified for the benefit of simplicity by Richard R. Reilly (1971). Darnell's method, called "clozentropy," is based on the theory of cloze testing in psychology and entropy in information theory. The last method uses multiple-choice scoring. It is slightly different from the usual multiple type tests in that the distructors are usually obtained by administering the open-ended version of the test to non-native speakers of English.

Several experiments have been conducted concerning the scoring methods of the cloze tests. Among them are Oller, John W. (1972), Stubb and Tucker (1974), Irvine, Atai and Oller (1974), Hinofotis, Frances B. (1976), and Brown, James D. (1978). Those studies indicate the following findings: first, the acceptable-word method is the most appropriate if the best overall cloze test of productive second/foreign language skill is desirable; and second, the correlation between the exact and acceptable word method is very high. Stubbs and Tucker (1974; 239-42) shows .97 and Irvine, Atai and Oller (1974; 249) also shows .94 in their studies of the correlation between the two.

2.2.2. Deleting Methods

Cloze tests use *n*th word deletion to be filled in the blanks by the examinees. It is said that you might use any deletion pattern.

It is, however, the general finding that a less-than-every-fourth-word, or more-than-every-tenth word deletion pattern is either unmanageable to take or impractical to construct (MacGinitie, W. H. 1961; 1121-1130).

According to the study by Alderson, J. Charles (1979; 224), it has been found that changing the deletion rate can have a drastic effect on the validity of the cloze test. It is the result of his experiment that "on an easy text, exact word scoring, changing the deletion rate from 6 to 8 results in a coefficient change of .59 to .70; and on the medium text, changing from rate 10 to 6 results in an increase in correlation from .57 to .86." It must be noticed, therefore, that the deletion rate variable produces a different kind of test even though you use the same text. It may lead to the fact that the deletion rate variable have a large amount of effect on the test validity and it produces the different test to measure different abilities of the examinees.

2.2.3. Length of the Text

Previous research has shown that it is recommended that you have about 50 items to be filled in in order to obtain sufficient information about the examinees' ability. For example, the minimum length of the words in the text, if you use the fifth-word deletion method, is about 250. If you use the seventh-word deletion method, about 350 words are needed.

In order to test this recommendation, Rand, Earl (1978; 62-71) has set up a program to see the effects of the test length and scoring method on the precision of cloze test scores. It is his finding that with twenty five items, the maximum reliability can be achieved across the four different scoring methods: exact-word, acceptable word, clozentropy and multiple-choice. And what he has concluded is that little precision is gained by making a cloze test longer than 25 items. This leads to the fact that if we have 25 blanks in the cloze test, we will be able to obtain sufficient information on the

ability of the examinees.

This finding helps a lot in improving the practicality of the cloze test. It is generally assumed that the longer the test is the higher reliability it obtains. It is very uneconomical, however, if we do not know what the minimum length of the test is. Rand's finding is of great use in that sense.

2.2.4. Difficulty Level of the Text

So far as the difficulty level is concerned, there has not been enough data in order to show the clear relationship between the level of difficulty and the validity of the test. As one of the conclusions, first of all, the writer would like to pick up what Oller (1979; 364) says. "True, some may be more difficult than others, but it has been demonstrated that for some purposes the levels of difficulty of the task does not greatly affect the spread of scores that will be produced."

It must be mentioned, however, that there is a tendency that difficult texts will result in better correlations with proficiency and criterion measures and that the text used might have an effect. In order to prove these tendencies, let us just look into what Alderson (1979; 222) concludes about the text variable. "There is a clear interaction between deletion rate and text which makes it impossible to generalize...Nevertheless, it is clear that different texts, using the same deletion rate, result in different correlations with the criterion, which suggests that different texts may well measure different aspects of EFL proficiency, or the same aspect more efficiently or less efficiently."

The two studies produce quite a different result, as you may easily understand. It is the writer's view that further study is needed. However, it is appropriate to say that the result of the later study is reasonable.

2.2.5. Content of the Text

It has often been suggested that one of the needed studies con-

cerns the content of the text used for the cloze test. It is regrettable to say that there has been no clear conclusion or findings concerning the content of the text.

It is, however, possible to state from the point of view of a general reading comprehension test of English as a foreign language that the subject matter should not be such as to give a marked advantage to students in particular fields. On the other hand, the text should not deal with information that is universally known. In that case, the students may be able to answer the question correctly without paying much attention to the content of the text.

Further study is sorely needed, even though the method of deciding on the best content of the text for the cloze test seems very difficult.

III. Problems

At the present stage, there are several problems of the cloze test pointed out by various researchers. The writer would like to concentrate his attention in this particular paper on three points. The first point the writer wants to pick up is the reliability of the cloze test. In the second place, the validity of the cloze test should be discussed. And lastly, the writer would like to discuss the factorial structure of the language proficiency in connection with the cloze test. Such discussion will help, the writer hopes, to assist readers in understanding the present status of the cloze test.

3.1. Reliability

3.1.1. The Reliability of the Cloze Test

Reliability, as one knows, refers to consistency of measurement and it is very important because of its relation to validity. A number of different ways of thinking about the concept have been derived in the field of educational psychology. There are, generally speaking,

three ways to estimate reliability: 1) test-retest method, 2) equivalent-forms method, and 3) tests of internal consistency. There are four basic methods in the tests of internal consistency: namely, Spearman Brown's formula, Kuder-Richardson formula, Cronbach's coefficient alpha, and Hoyt's analysis of varience procedure.

There has been considerable research on the reliability of the cloze test. The following studies indicate some of their findings:

Studies on Reliability of Cloze Test

Study	Sample size	Sample level	Number of items	Types of Reliability	Scoring method	Reliability coefficient
Darnell (1968)	48	Univ. (ESL)	50	Hoyt	CLOZEN	.86
Oller (1972)	398	Univ. (ESE)	50	K-R20	EX AC	.89–.92 .90–.95
Pike (1973)	430	Univ. (ESL)	25	K-R20	EX CLOZEN	.78–.91 .82–.85
Jonz (1975)	125	Grades 7–12	65	K-R20	MC	.95
Hinofotis (1976)	107	Univ. (ESL)	50	K-R20	EX AC	.61 .85
Brown (1980)	55 35	Univ. (ESL)	50	K-R20 test-retest	EX EX	.91 .81

CLOZEN=Clozentropy, EX=exact, AC=acceptable, MC=multiple-choice

It should be noticed that almost all the studies used the K-R20 formula in order to estimate the reliability of the cloze test. The formula 20 is:

$$r=\frac{k}{k-1}\left[1-\frac{\sum pq}{\sigma^2}\right]$$

where k is the number of items in the test, $\sum$ is the symbol for "the sum of", p is the proportion of correct response to particular item, q is the proportion of incorrect responses to that item (so that p plus q always equal 1), and σ^2 represents the variance of the scores on the test.

3.1.2. Independency of Relability

In connection with the use of K-R20, it should be noticed that it is applicable only to the test in which the items are scored by giving one point if answered correctly and nothing if not answered correctly. It should be also noticed that there is one fundamental assumption underlying almost any type of reliability coefficient, namely that the items consisting the test should be independent of one another. Let us consider one of the explanations concerning this point. Robert Ebel (1979; 275) says "The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained *independently* from the members of the same group." He continues, "...the operational definition calls for two or more *independent* measures, obtained from equivalent tests of the same trait for each member of the group."

If you study the cloze test carefully, one notices that the blanks you have to fill in, which are the test items, are contextually dependent on one another. The first blank is deeply connected with or related to the second one, which is again connected with or related to the third one. You have to create the word to be filled in, examining the context of the sentence. Therefore the items of the cloze test are said to be dependent on one another.

One may argue in the following way: If the cloze test is not providing reliable results just because of the lack of independency, is it not possible to criticize the reliability of the reading tests, listening tests, or even any kind of tests that measure overall internalized proficiency? Most of the reading or listening comprehension tests, one may continue, consist of a passage followed by several questions with multiple-choice type questions. Are the questions always independent of one another? Test items may be less dependent on one another than those of the cloze tests. There is, however, a great possibility of having dependency between items

because several question items are directly related to the passage or stem from which they are constructed.

The crucial point, in connection with these arguments, is that whether or not performance on one item influences (or is influenced by) the performance on the other items. Cloze tests or dictation tests have dependency among test items, while little or no dependency exist among test items in listening or reading comprehension tests if they are reliable. The reason for saying so can be verified by making several kinds of experiments. Suppose we have three forms in giving the cloze test. The first test is the normal cloze test. The second test is the one which is broken down into some parts. The third test is the one which is broken down further into still smaller parts. If the scores on the test have a tendency of decreasing as the passage is broken down into more independent segments, then we can say that the examinees take advantage of the context in answering the question. This will lead to the conclusion that cloze tests are contextually dependent. There have been some demonstrations that the examinee performance on a cloze test might depend on the context rather than independent items of the test. In that sense the following comment by Hossein Farhady (1979; 15) is worth mentioning: "Examining cloze and dictation types of tests reveals that they definitely violate the assumption of item independency. The items are contextually dependent on one another. Therefore, reliability coefficients, which are based on correlation between pairs of similar but independent items, will not be appropriately interpreted for cloze and dictation."

3.1.3. Alternative Method for Calculating Reliability

Are there alternative methods for calculating reliability which do not violate the assumption of item independency?

As stated before in this paper, there are three major ways to estimate such reliability: internal-consistency method, equivalent-forms method, and test-retest method. The internal-consistency

method (Split-half, K-R20, K-R21, or Cronbach alpha, etc) has been pointed out as one which violates the assumption of independence in the case of the cloze tests. Thus it probably presents an over-estimate.

Equivalent-form and test-retest methods can be used to circumvent the problem of item interdependency. There have been, however, a number of objections raised to these methods. In order to obtain equivalent-form reliability, we need parallel tests where equal variance, and equal correlations with any and all outside criteria. The construction of the parallel test that meets that criterion is very hard because of the nature of the cloze test. It may be said that the construction of parallel tests is almost impossible. Test-retest reliability also has a number of problems. We have therefore to be very careful in order to avoid these problems of teaching and learning effect when we use this method.

The writer tends to agree therefore with James D. Brown's statement at the Language Testing Conference 1980 at Albuquerque that "Test-retest reliability is a practical alternative to erroneous overestimates provided by internal-consistency estimates of reliability" in his paper "A Closer Look at Cloze".

3.2. Validity

3.2.1. The Validity of the Cloze Test

A test is valid to the extent that we know what it measures. In other words, we define the validity technically as the extent to which a test measures what it purports to measure. There are two basic approaches to estimate the degree of validity: logical analysis and empirical analysis.

Logical analysis includes content validity, item structure, and construct validity. Empirical analysis includes predictive validity, concurrent validity and construct validity. Since predictive and concurrent validation attempts to correlate performance on a measure

we are hoping to validate with an external criterion, they are also called criterion-related validity.

The following is the list of some of the studies on the validity of the cloze tests. It should be noticed that all of the validity were obtained by the concurrent validity. They are validated with an external criterion, such as Placement Test, TOEFL, etc.

Studies of Validity in Cloze Tests

Study	Sample size	Sample level	Criterion measure	Scoring method	Validity
Darnell (1968)	48	Univ. (ESL)	TOEFL	CLOZEN	.84
Oller (1972)	398	Univ. (ESL)	ESLPE, UCLA	EX AC	.73--.87 .80-.89
Ivine, Atai & Oller (1974)	159	Univ.	TOEFL	EX AC	.78 .79
Stubbs & Tucker (1974)	155	Univ. (ESL)	English Entrance Exam, Beirut	EX AC	.71 .76
Hinofotis (1976)	107	Univ. (ESL)	Placement Test CESL, SIU	EX AC	.80 .84
Hinofotis & Snow (1980)	66	Univ. (ESL)	Placement Test CESL, SIU	EX AC MC	.71 .74 .63

EX=exact, AC=acceptable, CLOZEN=clozentropy, MC=multiple-choice

3.2.2. What Does the Cloze Test Measure?

As seen in the list above, the validation of the cloze test has been carried out mostly from the point of view of concurrent validity. That is, research was conducted on whether cloze tests are correlated with concurrent criterion measures.

A series of correlational studies have found that the results of the cloze tests correlate highly with those of listening tests. Among them are the studies by Darnell (1968), Oller (1973a), Hinofotis (1976) (1980). Not only with the listening test results but also with the speaking test results, the researchers have shown that the results of the cloze test have a high correlation. Among those are the studies by Oller (1972) and Pike (1973). It has been their usual practice to

say that high correlation between cloze and listening or speaking leads to the fact that both are testing integrative performance. It has also been claimed that the pragmatic tests are superior to the discrete-point type of tests in that they tap the underlying, internalized expectancy grammar of the examinees.

There are several methods to estimate validity of tests. Concurrent validity is just one of them. The fundamental question concerning validity is whether or not the test is measuring what it is supposed to measure. What is it that it is supposed to measure in the cloze test? This may be the crucial point of the question. Is it the internalized expectancy grammar that the cloze test is supposed to measure? Is the cloze test constructed in order to measure the internalized expectancy grammar? Is it not possible to say that the internalized expectancy of grammar is not what the cloze test is supposed to measure, but that it is a means of explanation on the variance overlapping in the integrative or pragmatic tests? As Andrew D. Cohen (1980, 97) says, it may measure three types of knowledge: linguistic knowledge, texual knowledge, and knowledge of the world. The cloze test can be the measure of general reading comprehension. It can also be the measure of writing ability. So far as we know, the cloze test seems valid for the purpose of testing what language placement or entrance examination measure. In other words, the concurrent validity of the cloze test appears sound. It is, however, vague whether or not the cloze test measures "overall language proficiency." As Brown (1978; 20) and Farhady (1979; 13) say, "Because of the integrative nature of this type of test, it is not completely clear what the items are measuring." Furthermore, "It is a fact that no one has a clear idea of just what a cloze test is measuring."

The scond problem in connection with the validity of the cloze test concerns the face validity. Face validity refers to the acceptability of the test and test situation by the examinee or user. In

other words, what does the examinee think about filling in the blanks in the passage? Does he think that the man who can fill in the blanks has a good command of English? The examinees might think that the man who can fill in the blanks has a certain ability of vocabulary or grammar or even guessing, which is far from overall proficiency. The writer, as stated elsewhere (1978; 475), would like to point out that testing, whatever the form is, has curricular feedback. Even if some method seems to be able to measure real communicative proficiency, we cannot ignore the students' own assessment of how well the test reflect that particular proficiency. Face validity is extremely important in that very few students regard the people who can successfully fill in blanks on the paper as those who are necessarily able to use his or her overall proficiency.

3.2.3. Reliability and Validity

A test that is valid must be reliable, but a test that is reliable may or may be not be valid. In other words, you cannot have an unreliable and valid test, but you can have a reliable and invalid test.

The above statement is very important when we point out some of the problems in the cloze test. It seems very difficult to say that the cloze test is the best technique in order to evaluate the examinee's overall proficiency when we have a very close look at this test. As we have indicated, we have some problems in the reliability of the cloze test, and also in the validity of the cloze test.

As a tentative conclusion, the writer would like to quote from a passage from Hossein Farhady (1979; 17). "Based on the foregoing arguments, one may conclude that neither cloze nor dictation provides interpretable information on examinee performance. This may be due, in part, to the lack of validity (in its technical sense) and, in part, to the lack of reliability (in terms of violating the theoretical assumptions). Therefore, despite the fact that cloze and dictation can serve as a useful teaching devices, they may not be appropriate testing instruments. That is, they may not provide reliable or valid

results which enable administrators to make decisions on the basis of the scores on such tests."

3.3. Factorial Structure of Language Proficiency

3.3.1. Factor Analysis of the Test Result (1)

In order to verify whether language ability or proficiency consist of one factor or more, factor analysis is a powerful technique. Factor analysis refers, as is known, to the techniques for analyzing test scores in terms of some number of underlying factors. There have been a large number of research studies on the factor analysis of foreign language test results. It seems to the writer, however, that there are two different ways to understand the factor underlying the foreign language test results. One is the indivisibility hypothesis and the other is the divisibility hypothesis or the partial divisibility hypothesis.

The first hypothesis is supported by John W. Oller and his colleagues. One of the examples of their supports may be found in the result of the analysis of the data between 1969 and 1972 on UCLA's English as a Second Language Placement Examination. Oller (1979; 429) clearly states that "In other words, both the divisibility and partial divisibility hypotheses were clearly ruled out. Since all five of the test batteries investigated were administered to rather sizable samples of incoming foreign students the results were judged to be fairly conclusive. Whatever the separate grammar, reading, phonology, composition, dictation and cloze tests were measuring, they were apparently all measuring it. Some of them appeared to be better measures of the global language proficiency factor, but all appeared to be measures of that factor and not much else." The same conclusion that the results supported the indivisibility hypothesis was found in the study by Irvine, Atai, and Oller (1974), Oller and Hinofotis (1980) and Scholz *et al.* (1980). Here are the results of the factor analysis by Scholz *et al.* (1980; 32).

Varimax Rotated Solution for the Five Subscales of the FSI Oral Interview, the Three Subtests of the CESL Placement Test, and the Eighteen Subtests of the CESL Testing Project

(N=65 to 162 subjects)*

Test	Factor 1	Factor 2	Factor 3	Factor 4
CELT Listening Comrehension			.46	.56
Listening Cloze (Open-Ended)	.42	.44	.56	
Listening Cloze (Multiple-choice)				
Multiple-choice Listening Comprehension				
Dictation	.84			
Oral Interview-Accent			.72	
Oral Interview-Grammar		.85		
Oral Interview-Vocabulary		.83		
Oral Interview-Fluency		.76		
Oral Interview-Comprehension	.39	.81		
Repetition	.38	.34	.58	
Oral Cloze (Spoken Responses)	.46		.62	
Reading aloud	.34	.38		.49
CESL Reading				.80
Multiple-choice Reading Match	.74	.42		
Standard Cloze	.77			.41
Essay Ratings	.50	.43		.50
Essay Score	.63			
Multiple-Choice Writing	.63	.37		.41
Recall Rating	.43	.42		.63
CELT Structure				.74
Grammar (Parish Test)	.60			.51

* Only factor loadings above .32 ($p<.05$ with 65 df) are reported. The significant loadings on all four factors account for 57% of the total variance in all the tests.

As seen in the table above, all the scales of the FSI type Oral Interview, with the exception of the accent rating, load heavily on Factor 2. It is their another finding that the three subtests of the CESL placement test load mostly on Factor 4. However, the rest of the experimental tests are scattered all over the four factors in no discernible pattern. Scholz *et al.* (1980; 33) concludes that "...no clear pattern appears in which tests are grouped according to the

posited skills of listening, speaking, reading, and writing, or components of phonology, lexicon, or grammar, the data seem to fit best with the unitary competence hypothesis; and the divisible competence hypothesis is thus rejected."

3.3.2. Factor Analysis of the Test Results (2)

In spite of such strong support of the unitary competence hypothesis, it should be noticed that there is in fact another view; the divisible or the partial divisible hypothesis. Here is one of the examples of that support by Hossein Farhady in Testing Symposium, University of New Mexico, June 19-21, 1980.

Varimax Rotated Factor Loadings on the Functional Test and the ESLPE Subtests (N=416)

Subtest	Factor 1	Factor 2	Factor 3	Factor 4
Cloze	.47	—	.70	—
Dictation	.34	.60	.53	—
Listening Comprehension (Visual)	—	.76	—	—
Listening Comprehension (Written)	.35	.70	—	—
Reading Comprehension	.51	.33	.46	—
Grammar (Verbs)	.76	—	.32	—
Grammar (Prep.)	.73	.38	—	—
Grammar (Other)	.72	—	.33	—
Functional Test	.63	.36	.32	.32

One of the big findings in this analysis is that the data presented here seems to support that the unitary factor hypothesis is just questionable. The main reason for saying so is that each factor is loaded heayily from some particular subtests, not from most of the subtests. The data shows that there are heavy loadings on Factor 1 from Grammar (Verbs), (Prep.) and (Other). Factor 2 is loaded from Dictation, Listening Comprehension (Visual) and (Written). There are heavy loadings on Factor 3 from Cloze and Dictation. And Factor 4 is loaded mainly from the Functional Test.

Further data for supporting that the unitary factor hypothesis

is just questionable is the result obtained by the writer at UCLA under the guidance of Evelyn Hatch and Hossein Farhady. The subjects were 55 non-native speakers of English. The data analysis was conducted on June 9, 1980.

Varimax Rotated Factor Matrix

Subtest	Factor 1	Factor 2	Factor 3	Factor 4
Vocabulary	.64	.08	.25	.17
Grammar	.13	.15	.84	−.22
Reading Comprehension	.15	.11	−.13	.69
Cloze	.06	.51	.22	.05
Dictation	.20	.38	.27	.14
Listening (Visual)	.17	.69	−.07	.05
Listening (Written)	.76	.26	−.02	.07

The data shows that there are heavy loading on Factor 1 from Vocabulary and Written Listening Comprehension. Factor 2 is loaded from Cloze and Visual Listening Comprehension. Factor 3 is heavily loaded from Grammar. There are heavy loadings on Factor 4 from Reading Comprehension. As it is said that .30 and up is the cut-off criterion for the selection of the important factor, we may add that Factor 2 is loaded from Dictation, too.

The results obtained by Farhady and the writer seem to fit best with the assumption that the unitary competence hypothesis is just questionable. There seems to be some clear pattern in which tests are grouped according to the posited skills of listening, speaking, reading and writing, or components of phonology, lexicon, or grammar. However, the number of cases is so limited that it is necessary to have further research in order to say so definitely. If we could say with confidence that the unitary competence hypothesis is just questionable, the value of the logic underlying cloze tests should be reconsidered because of the unitary competence hypothesis in the logic underlying that test.

3.3.3. Meaning of Correlation Coefficient

There have been some questions as to why cloze tests concerned with written language skill should correlate so highly with spoken language skills such as dictation and listening comprehension. It was answered by Oller as stated before. Based on his assumption, the reason for the existence of such high correlation between them can be explained by "expectancy grammar" which is supposed to be the central component of language competence. The problem here the writer would like to refer to is whether or not highly correlated tests assess the same underlying factors in the same way. In other words, is it appropriate to say that the two highly correlated tests are measuring the same thing, i.e., expectancy grammar?

According to the theory developed by factor analysis, the total variance of a test could be regarded as the sum of three components: 1) common factor variance, 2) specific variance, and 3) error variance. Since the common variance may be made up of the combination of more than one common underlying factor, and since the sum of all variance components must equal 1, we have the following formula:

$$V_{\text{factor 1}}+V_{\text{factor 2}}+V_{\text{factor 3}}+V_{\text{factor 4}}+V_{\text{specific}}+V_{\text{error}}=1.00\ .$$

It is also found that the correlation coefficien between two tests is equal to the sum of the cross product of their common-factor loadings. Suppose we use that result of factor analysis. Then, what kind of interpretation can we obtain?

Test	Common factor			r_{AB}
	Factor 1	Factor 2	Factor 3	
A. cloze	.89	.25	.70	.83
B. listening	.35	.66	.50	.83
⋮	⋮	⋮	⋮	⋮

$r_{AB}=(.89)(.35)+(.25)(.66)+(.70)(.50)=.83$

The hypothetical data mentioned above says that there are heavy loadings on Factor 1 from cloze test. Factor 2 is loaded from listening test, and there are heavy loadings on Factor 3 from the cloze test. The correlation coefficient between the cloze and the listening test is thought to be high: .83. It should be, however, noticed that the proportion of these factors for each test is quite different. This hypothetical data implies that there is a case that even if the two tests correlate highly with each other, they do not test the same thing underlying factors in the same way. In other words, the tests with high correlation do not necessarily test the same thing such as grammar of expectancy or internalized grammar in the same way.

IV. Conclusion

We have had an overview of the underlying theory and major findings of the cloze test of English as a second or foreign language. And we have also discussed the problems concerning the reliability, validity and factor analysis of the test. As a conclusion of this short paper the writer would like to state that further research is needed in order to determine the real components of the foreign language proficiency. As seen in section **3.3.**, Factorial Structure of Language Proficiency, the technique of factor analysis may help a lot in order to deepen the research on that matter. The critical comments made by the writer may suggest that the direction of Oller's hypothesis is just questionable. The writer, however, has to delay his final conclusions until he has more convincing data of his own through experimentation.

Finally the writer would like to add that a very promising experimental study is underway at UCLA by Hossein Farhady, together with Evelyn Hatch, Frances Hinofotis and others, who are attempting to establish a test of communicative ability based on the concepts contained in the functional-notional syllabuses. The

writer is convinced that the findings from this exciting study will provide test specialists with a new direction in foreign language testing.

The writer also feels a debt of gratitude towards Evelyn Hatch, Frances Hinofotis, and Hossein Farhady of UCLA who contributed to the inspiration to write this paper. Special thanks are due to Mr. E. M. Carmichael of Kanagawa University for his assistance and suggestions.

References

Alderson, J. C. 1979. The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Q.* Vol. 13, No. 2, 212–227.

Brown, J. D. 1978. Correlational Study of Four Methods for Scoring Cloze Tests. Unpublished Master's Thesis, UCLA.

———. 1980. A Closer Look at Cloze. Paper presented at Language Testing Conference, Albuquerque, New Mexico.

Burt, M. K., H. C. Dulay, and E. Hernandez-Chavez. 1973. *The Bilingual Syntax Measure.* Harcourt-Brace and Jovanavich.

Carroll, J. B. 1961. Fundamental Considerations in Testing for English Language Proficiency of Foreign Students. In Center for Applied Linguistics *Testing the English Proficiency of Foreign Students,* CAL, 30–40.

———. 1968. The Psychology of Language Testing. In Davies, A. (Ed.) *Language Testing Symposium.* Oxford Univ. Press, 46–69.

Clark, J. L. D. 1972. *Foreign Language Testing: Theory and Practice.* Center for Curriculum Development.

Cohen, A. D. 1980. *Testing Language Ability in the Classroom.* Newbury House.

Darnell, D. K. 1968. The Development of an English Language Proficiency Test of Foreign Students Using a Clozentropy Procedure. Univ. of Colorado, US DHEW Project No. 7-H-010.

Ebel, R. 1979. *Essentials of Educational Measurement.* (Third Edition) Prentice-Hall.

Farhady, H. 1979. New Directions for ESL Proficiency Testing. Mimeo, UCLA.

Harris, D. P. 1969. *Testing English as a Second Language.* McGraw-Hill.

Hinofotis, F. B. 1976. An Invistigation of the Concurrent Validity of Cloze Testing as a Measure of Overall Proficiency in English as a Second

Language. Unpublished Ph. D. Dissertation, Southern Illinois Univ.
Hinofotis, F. B. and B. G. Snow. 1980. An Alternative Cloze Testing Procedure: Multiple-Choice Format. In Oller, J. W. and K. Perkins (Eds.) *Reseach in Language Testing*. Newbury House, 129–133.
Ilyin, D. 1976. *Ilyin Oral Interview*. Newbury House.
Irvine, P., P. Atai, and J. W. Oller. 1974. Cloze, Dictation, and the Test of English as a Foreign Language. In *Language Learning* Vol. 24, No. 2, 245–252.
Jakobovits, L. A. 1969. A Functional Approach to the Assessment of Language Skills. In *Journal of English as a Second Language*. 4, 63–76.
Jonz, J. 1975. Values of Hierarchies of Three Groups of Spanish-English Multilingual Adolescents. Doctoral dissertation, Univ. of New Mexico.
Lado, R. 1961. *Language Testing*. McGraw-Hill.
MacGinitie, W. H. 1961. Contextual Constraint in English Prose Paragraphs. *Journal of Psychology*, 51, 1121–1130.
Ohtomo, K. 1978. Testing Overall English Language Proficiency of Japanese Students. In Koike *et al.* (Eds.) *The Teaching of English in Japan*. Eichosha, 463–477.
Oller, J. W. 1972. Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language. *MLJ*. LVI (3), 151–158.
. 1973a. Cloze Tests of Second Language Proficiency and What They Measure. *Language Learning*. 23(1), 105–118.
. 1973b. Discrete-point Test versus Tests of Integrative Skills. In Oller, J. W. and J. C. Richards (Eds.) *Focus on the Learner*. Newbury House, 184–199.
. 1976. A Program for Language Testing Research. *Papers in Second Language Acquisition; Language Learning Special Issue* No. 4, 141–165.
. 1978. Pragmatics and Language Testing. In Solsky, B. (Ed.) *Advences in Language Testing: Series 2*, Center for Applied Linguistics, 39–57.
. 1979. *Language Tests at School: Progmatic Approach*. Longman.
Oller, J. W. and F. B. Hinofotis. 1980. Two Mutually Exclusive Hypotheses about Second Language Ability: Indivisible Competence. In Oller, J. W. and K. Perkins (Eds.) *Research in Language Testing*, Newbury House, 12–23.
Oller, J. W. and Perkins. (Eds.) 1980. *Research in Language Testing*. Newbury House.
Pike, L. W. 1973. An Evaluation of Present and Alternative Item Formats for Use in the Test of English as a Foreign Language. Unpublished ms. ETS.
Rand, E. 1978. The Effects of Test Length and Scoring Method on the Precision of Cloze Test Scores. *Workpapers in Teaching English as*

a Second Language. UCLA, Vol. XII, 62-71.

Reilly, R. 1971. A Note on 'Clozentropy: A Procedure for Testing English Language Proficiency of Foreign Students' *Speech Monographs*. 38(4), 350-353.

Scholz, G., D. Hendricks, R. Spurling, M. Johnson and L. Vanderburg. 1980. Is Language Ability Divisible or Unitary? A Factor Analysis of 22 English Language Proficiency Tests. In Oller, J. W. and K. Perkins (Eds.) *Research in Language Testing*, Newbury House, 24-33.

Spolsky, B., B. Sigurd, M. Sato, E. Walker, and C. Aterburn. 1968. Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency. *Language Learning*. Special Issue No. 3, 79-101.

Spolsky, B. 1978. Introduction: Linguists and Language Testers. In Spolsky, B. (Ed.) *Advances in Language Testing* Series 2, Center for Applied Linguistics, V-X.

Stubbs, J. B., and G. R. Tucker. 1974. The Cloze Test as a Measure of ESL Proficiency for Arab Students. *MLJ*, 58, 239-241.

Taylor, W. L. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly*. 30, 415-433.

———. 1954. Application of 'Cloze' and Entropy Measures to the Study of Contexual Constraints in Samples of Continuous Prose. Unpublished Ph. D. Dissertation, Univ. of Illinois.

———. 1956. Recent Development in the Use of 'Cloze Procedure' *Journalism Quarterly*. 33, 42-48.

Upshur, J. A. 1973. Productive Communication Testing: Progress Report. In Oller, J. W. and J. Richards (Eds.) *Focus on the Learner*, Newbury House, 177-183.

Valette, R. M. 1967. *Modern Language Testing: A Handbook*, Harcourt Brace.